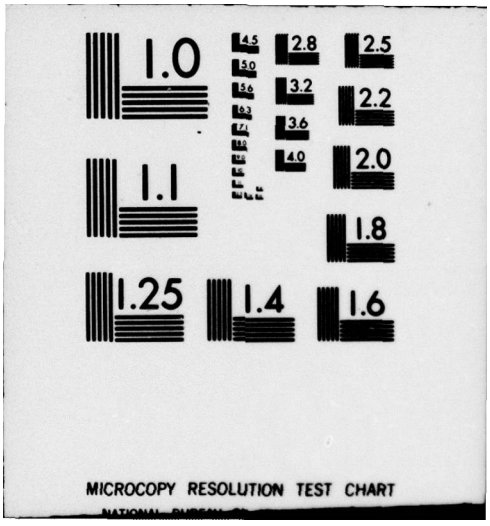
AD-A069 019 TEXAS UNIV AT AUSTIN CENTER FOR CYBERNETIC STUDIES F/G 9/4
INFORMATION THEORETIC ANALYSIS OF QUESTIONNAIRE DATA.(U)
MAR 79 P BROCKETT, P HAALAND, A LEVINE N00014-75-C-0616
UNCLASSIFIED CCS-336 NL

1 OF 1
AD
A069019

END
DATE FILMED
7-79
DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3
MICROCOPY RESOLUTION TEST CHART

LEVEL II (10) NW

CENTER FOR CYBERNETIC STUDIES

The University of Texas
Austin, Texas 78712

(10)

(9) Research Report (14) CCS-336

(6) INFORMATION THEORETIC ANALYSIS OF QUESTIONNAIRE DATA.

by

(10) P. Brockett*
P. Haaland**
A. Levine***

(12) 26 p.

(11) March 1979

This paper was presented in special session on Statistical and Information Theoretic Modeling, The Institute for Management Science/Operations Research Society of America joint meeting, New Orleans, Louisiana, April 30, 1979.

*Department of Mathematics and Department of General Business, The University of Texas at Austin, Austin, TX 78712

**Department of Mathematics, University of Miami, Coral Gables, FL 33124

***Department of Mathematics, Tulane University, New Orleans, LA 70118

(15) This research was partly supported by Project NR047-021, ONR Contracts N00014-75-C-0616 and N00014-75-C-0569 with the Center for Cybernetic Studies, The University of Texas at Austin. Reproduction in whole or in part is permitted for any purpose of the United States Government.

CENTER FOR CYBERNETIC STUDIES

A. Charnes, Director
Business-Economics Building, 203E
The University of Texas at Austin
Austin, TX 78712
(512) 471-1821

406 197

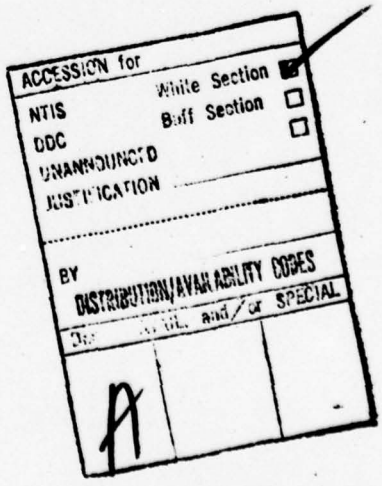
B

Abstract

We consider three important problems in the analysis of categorical questionnaire data. First, assessment of question worth and variable selection, second, the assessment of question validity using a pretest, and third, discrete discriminant analysis when the data is non-ordinal. The unifying approach used throughout is the concept of information theoretic distance measures. Simulations and applications to real data are presented.

1970 AMS Subject classification: Primary 62H30, 62L99, 62P99.

Key words and phrases: Categorical questionnaires, reliability, validity, discriminant analysis, variable selection, information divergence.



§1. Introduction

The analysis of categorical questionnaires poses many interesting problems of which we shall consider three: the assessment of which questions are worthwhile and which questions should be excluded (variable selection), the assessment of question validity and overall questionnaire validity, and the problem of discriminant analysis using categorical questionnaire data. These three problems are considered here as variants of a single problem which we shall attack using information theoretic techniques.

The use of information theoretic techniques is especially appealing in the analysis of questionnaire data since the entire purpose of such data is to answer some specific queries and the worth of each question should be determined according to how much information is supplied by the question towards answering these queries. To make this mathematically rigorous, suppose we wish to decide whether a respondent belongs in group 1 or group 2 with respective generalized densities f_1 and f_2 with respect to some measure λ. If the prior probabilities of group i membership are π_i, $i = 1,2$, then the log odds ratio in favor of group 1 membership is $\ell n\, \pi_1/\pi_2$. If an observation $\underset{\sim}{x}$ is made on the respondent, Bayes' Theorem may be used to determine the new posterior log odd ratio in favor of group 1 membership. The difference between the posterior and prior log odds ratio is taken as a measure of the amount of information supplied by the observation $\underset{\sim}{x}$ for discrimination in favor of group 1 membership. One easily works out that this difference as $\ell n(f_1(\underset{\sim}{x})/f_2(\underset{\sim}{x}))$ and this quantity is called the information gain from the observation in favor of group 1 membership, or simply the information gain (cf. Kullback (1959)). The expected information gain is obtained by randomizing $\underset{\sim}{x}$ according to the density f_1 obtaining the directed information measure

$$I(f_1|f_2) = \int \ell n \frac{f_1(x)}{f_2(x)} f_1(x)\lambda(dx) \ .$$

The symmetric measure of information between the two groups is called the divergence between the groups and is denoted by

$$J(f_1,f_2) = I(f_1|f_2) + I(f_2|f_1) = \int (f_1(x)-f_2(x)) \ell n \frac{f_1(x)}{f_2(x)} \lambda(dx) \ .$$

For the categorical questionnaires we shall be considering, we take λ as counting measure and the integrals become summations; $J(f_1,f_2) = \sum_i (p_i-q_i)\ell n(p_i/q_i)$ where $f_1(x_i) = P_1[X=x_i] = p_i$ and $f_2(x_i) = P_2[X=x_i] = q_i$.

§2. A measure of question validity

We assume that the purpose of the questionnaire is to obtain a summary index of how much of a certain attribute is possessed by the respondent. For example a psychiatric screening exam might measure how much "mental stress" is exhibited by a respondent, while in an industrial context, a quality control checklist might measure how much "propensity to fail" is exhibited by a certain machine. Employment screening exams which hope to measure a candidate's potential job success are another example.

A commom method of assessing the reliability and/or validity of a particular question in questionnaires such as those outlined above is to compare a respondent's overall questionnaire score with the score obtained on that particular question. The method we propose here is in this vein. We divide the respondents into quartiles, Q_1 , Q_2 , Q_3 and Q_4 , based upon their overall questionnaire scores excluding the question we wish to assess, and we measure the worth of that particular question by the amount of information it possesses for discriminating between these high and low scorers.

Let $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_k$ denote the proportion of high scorers (group Q_4) and $\hat{q}_1, \hat{q}_2, \ldots, \hat{q}_k$ denote the proportion of low scorers (Q_1) responding to the k answers to the question under consideration, and suppose there are n respondents in each of the reference groups Q_1 and Q_4. A measure of the amount of information in the question for discriminating between Q_1 and Q_4 (and hence a measure of the worth of the question) is given by taking a linear function of the estimated information theoretic divergence between Q_1 and Q_4. We define the D-value of the question to be

$$D = n/2 \sum_{i=1}^{k} (\hat{p}_i - \hat{q}_i) \ell n (\hat{p}_i / \hat{q}_i) .$$

We would discard a question if D is too close to zero indicating there is not sufficient information furnished by the question to discriminate between the high and low questionnaire scorers. Kullback (1959) shows that under a null hypothesis of $(p_1, \ldots, p_k) = (q_1, \ldots, q_k)$ (corresponding to the question having no discriminatory value), the asymptotic distribution of D is $\chi^2(k-1)$. Thus, our procedure is to retain a question only if $D > \chi^2_{1-\alpha}(k-1)$ where $\chi^2_{1-\alpha}(k-1)$ is the $1-\alpha$-th quantile of the $\chi^2(k-1)$ distribution. The probability of erroneously including a nondiscriminating question by using this procedure converges to α as the sample size increases. Another advantage of this procedure is that it should aid in establishing questionnaire validity since using this procedure includes only questions of proven discriminatory worth in the final questionnaire. For questionnaires such as employment screening questionnaires in which for legal reasons each question's inclusion must be justified, this method should be useful.

§3. Variable selection; which questions should be included

We shall again assume that the questionnaire is categorical, and we shall evaluate a question or sequence of questions by how much information is contained for

discriminating between two pre-given groups. (These may, of course, be Q_1 and Q_4 as in the previous section, however any two groups which we wish the questionnaire to distinguish will also serve our purposes.) We assume that we are given n_1 respondents from group one and n_2 respondents from group two, and we wish to develop a sequential procedure for determining which questions to include in the questionnaire analogous to the stepwise selection of variables in regression analysis.

For a particular question X, let J(X) denote the divergence between the probability distributions of group 1 and group 2 over the question. I.e.

$$J(X) = \sum_{x=1}^{k} (\hat{p}_x - \hat{q}_x) \ell n (\hat{p}_x / \hat{q}_x) \ .$$

It tells us the amount of information in question X for discriminating between the groups 1 and 2 with empirical response probabilities $\hat{p}$ and $\hat{q}$ respectively over the k_X answers to question X.

Our sequential procedure begins by choosing for first inclusion the most informative question X for discriminating. In this first step our procedure is similar in philosophy to that described by Levine (1974), Brockett, Haaland and Levine (1977b) and by Goldstein and Dillon (1977), (see also Goldstein and Dillon 1978), for selecting binary variables for inclusion in a multiway contengency table discrimination framework. In our case, however, we cannot assume that two categorical questions have the same number permissible categorical responses, e.g., the questions "Sex" and "Income level" may have markedly different number of response categories. This prohibits us from using the Goldstein-Dillon procedure. We shall use the quantity $D(X) = n_1 n_2 / (n_1 + n_2) J(X)$ as a measure of information in question X for discrimination (cf. Kullback (1959), Gokhale and Kullback (1978)). Asymptotically D(X) has a $\chi^2(k_X - 1)$ distribution, and this is why direct comparison of the cal-

culated $D(X)$ values is impossible. A question with $k_X = 11$ answers is expected to have a $D(X)$ value of 10 while a question with $k_X = 2$ would be expected to have a $D(X)$ value of 1 under the hypothesis that the question is not discriminating. This does not directly imply however that the question with 11 answers is more desirable than the question with 2 answers.

Since $D(X)$ has a $\chi^2(k_X-1)$ distribution under the null hypothesis that the two groups respond the same to the question, and has a non-central $\chi^2(k_X-1)$ distribution with a non-centrality parameter equal to the discriminatory power of the question in the case where the alternative hypothesis holds and the question actually discriminates, we shall use instead the p-value of the $D(X)$ statistic as a measure of discriminatory power of question X. If $p_X = P[\chi^2(k_X-1) \geq d]$ where d is the observed value for $D(X)$, then the smaller p_X, the more informative is question X. Although the values $D(X)$ for various questions X may not be directly comparable in general (as they would be for example if k_X was always the same), the p-values p_X are always comparable and easily calculated from readily available χ^2 tables. (Alternately, if no tables are available, the normalized quantities $Z_X = \frac{D(X)-(k_x-1)}{\sqrt{2(k_x-1)}}$ would be quickly comparable variables, approximately normal zero-one for k_x large.)

Using the p-values, which are distributed uniformly over $[0,1]$ under the null hypothesis of no discriminatory power, we select as the first question that question X with minimum p_X value, provided this p_X value is significantly small. We can assess significance for $\min_{1 \leq X \leq m} p_X = U_{(1)}$ by using the distribution function $F(t) = 1 - (1-t)^m$ for $0 \leq t \leq 1$ as the c.d.f. for $\min_{1 \leq X \leq m} p_X$. Thus the best question has significant discriminatory power at level of significance α if $\min_{1 \leq X \leq m} p_X \leq 1 - (1-\alpha)^{1/m}$. (The Goldstein-Dillon procedure does not employ the actual

distribution for their selection statistic, and hence will not lead to a fixed type 1 error.) Having chosen the first question for inclusion according to this procedure, we select the second question for inclusion as that question which yields the maximum additional information to the already selected first question. For notational convenience, relabel the questions so that the first question selected is called question 1. We look at all question pairs $(1,Y)$, $2 \leq Y \leq m$ and consider the joint probabilities p_{xy} and q_{xy} for the two groups over the possible answer pairs (x,y) on questions $(1,Y)$.

The quantity $D(1,Y) = n_1 n_2/(n_1+n_2) \sum_{x=1}^{k_1} \sum_{y=1}^{k_Y} (p_{xy}-q_{xy}) \ell n\, p_{xy}/q_{xy}$ is a measure of the joint discriminatory power of the question pair $(1,Y)$, and hence $D(1,Y) - D(1)$ is a measure of the increase in discriminatory information obtained by adding question Y. Note that $D(1,Y) - D(1) = n_1 n_2/(n_1+n_2) \sum_x \sum_y (p_{xy}-q_{xy}) \ell n (p_{xy}/q_{xy})$ $- n_1 n_2/(n_1+n_2) \sum_x (p_x - q_x) \ell n (p_x/q_x) = n_1 n_2/(n_1+n_2) \sum_x \sum_y (p_{xy}-q_{xy}) \ell n (p_{xy} q_x / p_x q_{xy})$ $= n_1 n_2/(n_1+n_2) \sum_x p_x I(p_{Y|x} | q_{Y|x}) + n_1 n_2/(n_1+n_2) \sum_x q_x I(q_{Y|x} | p_{Y|x})$ where $p_{Y|x}$ and $q_{Y|x}$ are the conditional probability distributions of groups 1 and 2 respectively over the answers to question Y given that question 1 was answered x, i.e., $p_{Y|x}(y) = p_{xy}/p_x$. This equality implies $D(1,Y) - D(1) \geq 0$ with equality only if Y contains no additional information given the answer to question 1 (i.e. the addition of Y can only improve things). This equation also shows that the distribution of $D(1,Y) - D(1)$ is a weighted sum of (non-independent) χ^2 variables, the weights reflecting the probability of a particular response x to question 1, and the χ^2 variable measuring the information expected to be added by question Y given that particular response x to question 1. A stepwise regression analogue would consider $\{D(1,Y)-D(1)\}/D(1)$ as a measure of increased discriminatory power obtained by the addition of question Y. The distributional properties of this ratio have not been

explored in full, however when the question responses are independent variates, one has $D(1,Y) = D(1) + D(Y)$ so the above ratio has (asymptotically) an F distribution with parameters (k_Y-1, k_1-1) (see Brockett, Haaland and Levine (1977a,b)). When the responses are not independent, this procedure is conservative in that the true ratio is stochastically dominated by the given F distribution.

In this paper we shall also present a different approach based upon information theoretic analysis similar to that used in contingency table analysis (cf. Gokhale and Kullback (1978) and Kullback (1959)). This method is more convenient to use than Goldstein and Dillon's technique since one utilizes the entire set of answers to the previous question for determining the usefulness of a proposed new question rather than having to condition individually upon each possible response. Goldstein and Dillon's technique would result in different respondents obtaining different sequences of questions and perhaps different questions. Since the order of presentation of questions has been shown to make a statistically significant difference in questionnaire score (cf. Payne (1951), Oppenheim (1966), and for an up-to-date bibliography and study, see Kalton, Collins and Brook (1978)). We desire a unified approach to variable selection not dependent upon the particular answers given to previous questions of the Goldstein-Dillon method is useful for expensive binary medical tests.

Let x_{ijy} denote the number of respondents in group i $(i = 1,2)$ who answer j to question 1 and answer y to question Y, and $p(i,j,y)$ represent the corresponding proportion of respondents in group i with these answers. A test of the hypothesis that the inclusion of question Y yields no additional discriminatory power can be obtained by testing the hypothesis that the conditional distribution of Y is independent of the group classification given the answer to question 1, i.e.

$$H_{1,Y}^{(0)}: \; p(iy|j) = p(i|j)p(y|j) \; .$$

The worth of question Y is assessed by the p-value for rejecting $H_{1,Y}^{(0)}$. Using the directed divergence distance measure to test $H_{1,Y}^{(0)}$ yields the statistic

$$I(Y|1) = 2I(p(ijy)|p_{(i|j)}p_{(y|j)}p_{(j)}) = 2\sum_{i=1}^{2}\sum_{j=1}^{k_1}\sum_{y=1}^{k_Y} x_{ijy}\ell n\left(\frac{x_{ijy}x_{\cdot j\cdot}}{x_{ij\cdot}x_{\cdot jy}}\right)$$

where the dot replacing a subscript is the usual convention for "sum over that variable", i.e. $x_{ij\cdot} = \sum_{y=1}^{k_Y} x_{ijy}$ etc. The distribution of $I(Y|1)$ under $H_{1,Y}^{(0)}$ is asymotically χ^2 with $k_1(k_Y-1)$ degrees of freedom (c.f. Kullback 1959, or Gokhale and Kullback 1978). Hence if $p_Y = P[\chi^2(k,k_Y-k_1) \geq i(Y|1)]$ where $i(Y|1)$ is the observed value for $I(Y|1)$, we select the second question to minimize p_Y, $2 \leq Y \leq m$. The exact level of significance can be found from the distribution function $F(t) = 1 - (1-t)^{m-1}$, $0 \leq t \leq 1$.

If the p-value for the second question is significant, we proceed on to select the third question by the same procedure. We test if the group classification and the answer to question Z, $3 \leq Z \leq m$ are conditionally independent given the response to questions 1 and 2. The information statistic is

$$2\sum_{i=1}^{2}\sum_{j=1}^{k_1}\sum_{k=1}^{k_2}\sum_{\ell=1}^{k_Z} x_{ijk\ell}\ell n\left(\frac{x_{ijk\ell}x_{\cdot jk\cdot}}{x_{ijk\cdot}x_{\cdot jk\ell}}\right)$$

which is χ^2 (with $k_1k_2(k_Z-1)$ degrees of freedom. Again, the p-values for each question $3 \leq Z \leq m$ are compared to determine if the minimum p-value question is significant. If it is, we include it as question 3 and proceed until we obtain a non-significant result. When we finally find a non-significant result, we quit adding questions and consider the questionnaire complete. This procedure can reduce the size of the overall questionnaire.

One problem with utilizing contingency table methods of analysis is the rapid proliferation of cells, resulting in possible empty cells. Empty cells here won't

bother us, however zero marginals will. The problem of zero marginals, and the resulting loss of degrees of freedom is discussed in Gokhale and Kullback (1978). Another approach is to break the questionnaire into subsets of 3 or 4 questions each, *each group of questions being treated separately.* For example, a group of questions concerning economic status might be treated separately from a group concerning health, which in turn is treated separately from a group concerning education level. Each group is analyzed to obtain the questions in that particular group which should be included. The best subset of each group is then combined to form the overall questionnaire. Still another approach to the sparse cells problem is the "nearest neighbor" approach of Hills (1966). One groups together respondents whose previous answers differ in only one place on one question.

§4. Discrimination using information gain

The problem of discriminant analysis using categorical data has attracted much attention in the recent literature (c.f. Lachenbruch 1975, Goldstein and Dillon 1978 and the bibliographies contained therein). If the number of variables is quite small a contingency table approach is possible. For moderate numbers of variables, a log linear model may be fitted assuming certain higher order interaction terms vanish (c.f. Gokhale and Kullback 1978, and for computational techniques, see Brockett, Charnes and Cooper 1979). If the questions all have ranked responses, or perhaps binary responses, then Fisher's linear discriminant function (LDF) (Fisher 1936) has proven to be quite effective for classifying respondents into their correct group (c.f. Lachenbruch 1975). Fisher's LDF does not perform well, however, when the answers are not of a ranked character. Moore (1973) refers to this as reversal in the likelihood ratio.

Discriminant analysis for discrete data involves two processes; first one must score the categories, and second, one must combine the individual question scores to obtain an overall questionnaire score to be used for classification. The procedure

we shall discuss here involves information theoretic scoring. This method is effective for non-ordinal data, and outperforms Fisher's LDF when reversals are present.

Let $\underset{\sim}{p}_t = (p_{t1}, \ldots, p_{tk_t})$ and $\underset{\sim}{q}_t = (q_{t1}, \ldots, q_{tk_t})$ denote the group 1 and group 2 response probabilities for the k_t answers to question t. By scoring the i-th answer via the information gain $\ell n \frac{p_{ti}}{q_{ti}}$ in favor of group 1 membership, we may transform the non-ordinal data into ordinal data. The larger the score, the more likely is group 1 membership. (This is not true with raw scoring if the two groups' responses are polarized with respect to each other.) For simplicity we shall assume conditional independence of the responses given the group. This is a common assumption in medical diagnostics (c.f. Warner et al (1961), Bishop and Warner (1969), Boyle et al (1966), Nugent et al (1964) or Reale et al (1968)), but could be modified if obviously necessary by scoring subcollections of non-independent questions separately and then adding together the component subcollection scores to obtain an overall questionnaire score, or perhaps utilizing LDF for ordinal data, and our method for the non-ordinal questions. We call our method discrimination using information gain (DIG).

If we are given samples of size n_1 and n_2 from group 1 and group 2, we may estimate p_{ti} and q_{ti} from these training samples, and pick a number s^* such that $\varepsilon(1,2) + \varepsilon(2,1)$ is minimized, where $\varepsilon(1,2)$ is the percentage of the n_1 respondents in group 1 with score $< s^*$ and $\varepsilon(2,1)$ is the percentage of the n_2 respondents in group 2 with score $> s^*$. We classify a respondent into group 1 if his questionnaire score is $\geq s^*$ and into group 2 otherwise, so $\varepsilon(i,j)$ represent the percentage misclassified as belonging to group j.

Simulation studies were run to assess the power of this procedure relative to Fisher's LDF and the discriminant procedure available in the SPSS (Statistical Package for the Social Sciences) computer package. Also compared were the question weighting schemes of RAO (1970), SPSS (c.f. Cooley and Lohnes 1971), and the divergence weights from §2.

The first questionnaire we simulated is presented in Table 1. Each question was simulated independently according to the given probability structure. Note that questions 1,2,3 discriminate well but in a different way than do 5 and 10. The remaining questions discriminate moderately well except for 6 which essentially does not discriminate the groups.

Table 1 : Probability Structure of the simulated questionnaire I

Question No.	Probability of response of group 1 to part					Probability of response of group 2 to part				
	I	II	III	IV	V	I	II	III	IV	V
1	25	10	30	10	25	10	35	10	35	10
2	35	10	10	10	35	10	35	10	35	10
3	35	10	10	10	35	10	25	30	25	10
4	25	10	30	10	25	10	25	30	25	10
5	25	25	25	10	15	15	10	25	25	25
6	15	25	25	25	10	10	25	25	25	15
7	25	25	30	10	10	10	10	30	25	25
8	10	35	35	10	10	10	10	35	35	10
9	35	10	35	10	10	10	10	35	10	35
10	35	35	10	10	10	10	10	10	35	35

Table 2 shows the errors of misclassifications for a simulation run with 100 members in each group.

Table 2: $\varepsilon(i,j)$. Error of Misclassification in % of Group i into group j for SPSS, Fisher, DIG: 100 samples from each group.

Method Error (%)	Fisher	SPSS	DIG
$\varepsilon(1,2)$	26	21	1
$\varepsilon(2,1)$	18	17	3

As predicted, DIG does much better for this type of questionnaire since DIG does not depend upon the centroid separation of the two groups, but rather the "information gain" available from the questionnaire. This superiority of DIG also holds for the determination of significant questions. As shown in Table 3 the Rao and SPSS methods distort the relative discriminatory power of the questions while the DIG method correctly orders them according to the information present in the question for discrimination between the groups. The worth of a question which discriminates between the groups in a non-ordinal way is assesed by DIG but not the other two. This is because the DIG procedure measures the "distance" between the two probability measures corresponding to the two groups, and not the Euclidean distance between two real numbers (centroids) which ostensibly represent the groups.

Table 3: Question Weights for Three Methods.

Question / Method	1	2	3	4	5	6	7	8	9	10
Rao $F(1,191)$	.406	.052	.552	.094	3.92	0	13.8	8	5.5	38.8
SPSS $F(1,198)$	.025	.164	.909	.5	.533	.208	39.5	7.98	19.5	91.1
DIG X^2 (4)	44.3	56.1	81.5	14.7	12.9	5	43.5	32	19.4	83.2

We also ran the simulation for 10,000 samples in each group. For DIG, we found $\varepsilon(1,2) = 7.6\%$, $\varepsilon(2,1) = 5.6\%$, while for SPSS, $e(1,2) = \varepsilon(2,1) = 20\%$. These results verify the theoretical conclusion that for this type of questionnaire, with discriminating but non-ordinal questions, DIG is perferred. Examples of such questions might be marital status, or questions on political extremes.

Table 4 shows a more extreme example of a questionnaire which discriminates very well, but upon which the SPSS and Fisher methods will produce distorted results.

Table 4: Probability Structure of the simulated questionnaire II

Question No.	Probabilities of response of group 1 to part					Probabilities of group 2 to part				
	I	II	III	IV	V	I	II	III	IV	V
1	25	10	30	10	25	10	35	10	35	10
2	35	10	10	10	35	10	35	10	35	10
3	35	10	10	10	35	10	25	30	25	10
4	25	10	30	10	25	10	25	30	25	10
5	10	25	30	25	10	10	10	60	10	10
6	10	35	10	35	10	10	10	60	10	10
7	20	20	20	20	20	10	10	60	10	10
8	35	10	10	10	35	10	10	60	10	10
9	35	10	10	10	35	10	20	40	20	10
10	35	10	10	10	35	10	30	20	30	10

Table 5 shows the values of $\varepsilon(i,j)$ and Table 6 shows the values of the question "weights". Note that the SPSS and Fisher methods distort the classification and the determination of important discriminatory questions. This is because the centroid methods cannot distinguish the groups when they have this symmetric response pattern. Nevertheless the two probability distributions are some "distance" apart.

Table 5: $\varepsilon(i,j)$. Error of Misclassification in % of Group i into Group j for three methods: 100 samples in each group. Simulation II

		Fisher	SPSS	DIG
Error in %	$\varepsilon(1,2)$	44	41	1
	$\varepsilon(2,1)$	40	36	3

Table 6: Question Weights for three Methods: Simulation II

	Rao (F(1,194)	SPSS F(1,198)	DIG ($X^2(4)$)
1	.182	.0247	44.26
2	.134	.1643	56,08
3	6.444	2.909	81.45
4	.168	.4995	14.7
5	.058	.0396	28.1
6	.356	.0732	66.57
7	.002	.0035	44.87
8	.001	.0218	110.13
9	.76	.2262	69.33
10	2.49	1.1069	77.11

The simulations were run again using 10,000 samples in each group. For DIG, $\varepsilon(1,2) = 5\%$, $\varepsilon(2,1) = 4\%$; for SPSS, $\varepsilon(i,j) = 49\%$. The result again indicates that SPSS will be less reliable than DIG. The DIG weights are also more reliable indicators of question discriminatory power.

§6. Applications

We shall present several examples in which the information theoretic methods of analysis were used on actual data.

We consider first a psychiatric screening questionnaire developed by Dr. H. Davidian, Head of the Department of Psychiatry, University of Tehran, Iran. In this questionnaire the permitted responses to questions such as "Do you feel restless?"

are "never", "occasionally", "frequently", "always". The questionnaire was designed so that each question measures the degree of some aspect of the respondent's "mental stress"; the question responses are all ordered by degree in the same direction; "mentally ill" patients are presumed to respond to the "high" end, "normals" at the "low' end. Since this is so, the two group's score centroids should be well separated and consequently, both SPSS and DIG should discriminate well.

The questionnaire consisted of 46 questions and was designed for the purpose of classifying each respondent as "mentally ill" or "normal". It was given to 143 respondents, 90 of whom were classified prior to the administration of the questionnaire as "mentally ill" while the rest were "normal". The values assigned to the responses were 0 for "never", 1 for "occasionally", 2 for "frequently" and 3 for "always". SPSS was run using this raw scoring technique. Essentially SPSS and DIG behave the same for this nice ordinal data. The question weights developed by SPSS and DIG, as expected, gave essentially the same assessment of question worth. The question weights were used to shorten the questionnaire as outlined in §4. The twenty-two questions with the highest weights were selected. (Psychiatric technical considerations also played a part in the choice of these questions.) Using this new "reduced" questionnaire it was found that $\varepsilon(1,2) = \varepsilon(2,1) = 0$ for the DIG analysis, while for the SPSS analysis $\varepsilon(1,2) = 0$ and $\varepsilon(2,1) = 4\%$. This questionnaire has been used for screening purposes in Iran. The use of thr reduced questionnaire has resulted in a considerable saving in time over the original questionnaire. (Note: in all of these simulations the apparent error rate is used, and hence is optimistically biased. Still the results are encouraging.)

The second set of data upon which these methods have been used involved a survey conducted by the Pan American Health Organization (PAHO) on child mortality in 1969-1970 in South American countries. Among the questionnaires was one covering the socio-economic status of a household and various environmental factors. Some questions were "What type of water supply do you have?" with such permitted responses as "piped water", "well", "rain water"; "What is the marital status of the mother?", "married",

"divorced", "separated". We note here that many of these questions had answers which were not essentially ordinal and hence reversals may occur, and a linear discriminant function may be inappropriate. (Because of a contractual agreement between PAHO and the World Health Organization, Geneva, where this data was analyzed by one of us (A.L.) we are not able to disclose details of this questionnaire or of the analyses.) We analyzed twelve questions from this questionnaire. Group 1 consisted of all those households where a child under 5 died of malnutrition or diarrhea. The second group consisted of households where a child died from other causes. In all there were 952 households, however, complete information was available only on 154 in group 1, 37 in group 2. We chose only those questionnaires without missing responses since SPSS needs a special program for missing data (DIG does not) and we wanted a direct comparison between the two methods. The proportion of answers of each group to each response is given in Table 7. From Table 8 we find that for DIG, $\varepsilon(1,2) = 12$, $\varepsilon(2,1) = 24$ and for SPSS $\varepsilon(1,2) = 26$, $\varepsilon(2,1) = 25$. It should be noted that the results of SPSS and DIG agree upon which of the questions are significantly discriminating except for question 9 which DIG found to be highly discriminating and SPSS found to be not discriminating. The importance of question 9 was lost to SPSS since it discriminated in a non-ordinal manner. Consequently, if one uses SPSS on data which is not essentially ordinal, one may unintentionally eliminate significant variables.

Table 7: Simulation of response probabilities for Pan American Health Organization survey

Question No.	Probability of response of group 1 to part					Probability of response of group 2 to part				
	I	II	III	IV	V	I	II	III	IV	V
1	17	37	46	0	0	0	29	66	5	0
2	76	22	2	0	0	55	50	5	0	0
3	69	25	6	0	0	47	45	8	0	0
4	93	7	0	0	0	83	13	0	0	0
5	19	16	65	0	0	10	24	66	0	0
6	39	15	23	23	0	32	18	16	34	0
7	24	15	22	26	13	11	34	21	18	16
8	92	8	0	0	0	99	1	0	0	0
9	32	17	14	12	25	16	28	19	13	24
10	70	1	29	0	0	53	0	47	0	0
11	68	29	3	0	0	45	50	5	0	0
12	63	37	0	0	0	76	24	0	0	0

Table 8: Error (%) of Misclassification: Simulation III
100 samples in each group.

Method / Error	Fisher	SPSS	DIG
$\varepsilon(1,2)$	31	26	12
$\varepsilon(2,1)$	32	25	24

Table 9: Question Weights: Simulation III

Method \ Question	1	2	3	4	5	6	7	8	9	10	11	12
Rao	12.6	4.48	0	4.9	1.89	.48	.128	1.39	.808	2.41	11.5	.09
SPSS	24.9	10.6	.966	7.96	2.82	.929	.07	4.78	.106	3.95	27.7	2.82
DIG	34.3	10.8	4.85	8.45	4.58	4.45	8.11	5.52	11.2	3.95	28.1	2.82

As a final example of how this method has been used we briefly sketch the following: the acceptability group in the human reproduction division of the World Health Organization has used the DIG technique to assess the feasibility and acceptability of various modes of contraception. In particular they used this method to determine for which groups of people a paper birth control pill is acceptable. Various factors affect the acceptability of the paper pill. For example, the persons involved may refuse to eat paper, the climate may be such that the paper cannot be kept dry, or the life style may be such that the paper sheets cannot be kept clean. On the other hand if the paper pill is acceptable in a particular region, it reduces costs and is easier to store and administer. A categorical questionnaire was designed to ascertain which groups would accept the

paper pill and was given to samples in Alexandria and Cario (in Egypt), Cariche and Ibaden (India), Manilla (Philippines), Stockholm (Sweden) and Bangkok (Thailand). It was desired to find which questions or factors discriminated between those respondents who accepted the paper pill and those who did not. Also it was desirable to know how effective each question was in distinguishing between the groups.

Both SPSS and DIG analysis was performed on the data using sample sizes of about 200 in each country. Both produced the same set of discriminating questions, and approximately the same error rates (20-30%). We cannot present a detailed description of the data collected since this study is still ongoing, and the WHO has priority on the publication of the exact data. Nevertheless, this example and previous examples show how information theory has been successfully applied to real data. For further information on the contraceptive study, contact Dr. Cri Kars, Human Reproduction Division, WHO, Geneva, Switzerland.

There is also a user's manual being produced at the WHO by Busca and Diethelm which contains a computer package to implement the DIG analysis.

References

1. Bishop, C.R. and H.R. Warner (1969): "A Mathematical Approach to Medical Diagnosis: Application to Polycythemic States Utilizing Clinical Findings with Values Continuously Distributed", Computers and Biomedical Research 2, 486-493.

2. Boyle, J.A., W.R. Grieg, D.A. Franklin, R.M. Harden, W.W. Buchanan and E.M. McGirr (1966): "Construction of a Model for Computer-assisted Diagnosis: Application to the Problem of Non-toxic Goitre", Quarterly Journal of Medicine, N.S. 35, 565-588.

3. Brockett, P.L., A. Charnes and W.W. Cooper (1979): "MDI estimation via unconstrained convex programming", Center for Cybernetic Studies Report CCS 326, The University of Texas at Austin.

4. Brockett, P.L., P. Haaland and A. Levine (1977a): "Discriminant Analysis for Categorical Questionnaire Data", Tulane University mathematics department preprint.

5. Brockett, P.L., P. Haaland and A. Levine (1977b): "A characterization of divergence with applications to Questionnaire Information", to appear, Information and Control.

6. Busca, B. and P. Diethelm (1978): "Discriminant Analysis using Information Gain (DIG)", manual for WHO, in preparation.

7. Cooley, W.W. and P.R. Lohnes (1971): Multivariate Data Analysis, John Wiley and Sons, New York, Chapter 9.

8. Fisher, R.A. (1936): "The use of multiple measurements in taxonomic problems", Ann. Eugen., 7, 179.

9. Gokhale, D.V. and S. Kullback (1978): The Information in Contingency Tables, New York, Marcel Dekker, Inc.

10. Goldstein, M. and W.R. Dillon (1977): "A stepwise discrete variable selection procedure", Comm. Stat., Theory and Methods 6, 1423-1436.

11. _____(1978): Discrete Discriminant Analysis, New York, John Wiley and Sons.

12. Hills, M. (1966): "Allocation rules and their error rates", J. Roy. Stat. Soc. B 28, 1.

13. Kalton, Graham, M. Collins and L. Brook (1978): "Experiments in Wording Opinion Questions", Applied Stat., 27, No. 2, 149-161.

14. Kullback, S. (1959): Information Theory and Statistics, New York, John Wiley and Sons, Dover Press (1968), New York.

15. Lachenbruch, P.A. (1975): Discriminant Analysis, Hafner Press, New York.

16. Levine, A. (1974): "A new approach to Discriminant Analysis in Screening Questionnaires", Int. Symp. on Epidemiological Studies in Psychiatry, Tehran.

17. Moore, D.H. II (1973): "Evaluation of five discrimination procedures for binary variables", J. Ann. Stat. Assoc., 68, 339-404.

18. Nugent, C.A., H.R. Warner, J.T. Dunn and F.H. Tyler (1964): "Probability Theory in the Diagnosis of Cushing's Syndrome", The Journal of Clinical Endocrinology 24, 621-627.

19. Oppenheim, A.N. (1966): Questionnaire Design and Attitude Measurement, New York, Basic Books, Inc.

20. Payne, S.L. (1951): The Art of Asking Questions, Princeton University Press.

21. Reale, A., G.A. Maccacaro, E. Rocca, S. D'Intino, P.A. Geoffre, A. Vestri and M. Motolese (1968): "Computer Diagnosis of Congenital Heart Disease", Computers and Biomedical Research, 1, 533-549.

22. Rao, C.R. (1970): "Inference in discriminant function coefficients", Essays in Probability and Statistics, R.C. Bose, etal, eds., Chapel Hill, University of North Carolina and Statistical Publishing Society, pg. 487-602.

23. Warner, H.R., A.F. Toronto, L.G. Veasey and R. Stephenson (1961): "A Mathematical Approach to Medical Diagnosis: Application of Congenital Heart Disease", Journal of the American Medical Association, 177, 177-183.

Unclassified

Security Classification

DOCUMENT CONTROL DATA - R & D

Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)	2a. REPORT SECURITY CLASSIFICATION
Center for Cybernetic Studies The University of Texas at Austin	Unclassified
	2b. GROUP

3. REPORT TITLE

Information Theoretic Analysis of Questionnaire Data

4. DESCRIPTIVE NOTES (*Type of report and inclusive dates*)

5. AUTHOR(S) (*First name, middle initial, last name*)

P. Brockett, P. Haaland, A. Levine

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
March 1979	23	23

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
N00014-75-C-0569 & 0616	Center for Cybernetic Studies Research Report CCS 336
b. PROJECT NO NR047-021	
c.	9b. OTHER REPORT NO(S) (*Any other numbers that may be assigned this report*)
d.	

10. DISTRIBUTION STATEMENT

This document has been approved for public release and sale, its distribution is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Office of Naval Research (Code 434) Washington, DC

13. ABSTRACT

We consider three important problems in the analysis of categorical questionnaire data. First, assessment of question worth and variable selection, second, the assessment of question validity using a pretest, and third, discrete disciminant analysis when the data is non-ordinal. The unifying approach used throughout is the concept of information theoretic distance measures. Simulations and appplications to real data are presented.

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6811

Unclassified

Security Classification

A-31408

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Categorical questionnaires						
Reliability						
Validity						
Discriminant analysis						
Variable selection						
Information divergence						